Motor Trend - Effects of Transmisson on MPG

Douglas Wirtz

March 29, 2016

Introduction and Executive Summary

This data anaylsis was performed to specifications required by the course project of the Regression Models course in the Data Science Specialization offered by Johns Hopkins University on Coursera. In this project, the R dataset, `mtcars`, was used to explore the relationship between a set of variables and miles per gallon (`mpg`).

The data was used to answer the following questions:

- Is automatic or manual transmission better for MPG?
- What is the MPG difference between automatic and manual transmissions?

Based on this regression analysis, we can conclude that manual transmission cars have a higher miles per gallon (`mpg`). After ajusting the miles per gallon for horsepower (`hp`), number of cylinders (`cyl`), and weight (`wt`), manual transmission cars get 1.8 more miles per gallon when comparing the automatic transmission cars. To further the analysis, we can conclude that MPGs will decrease by 2.5 for every 1000-pound increase in weight. MPGs will decrease 0.32 for every 10 increase in horsepower. When the number of cylinders increase from 4 to 6, the MPGs decrease by 3. And finally, the MPGs will decrease by another 2.2 when the numbers of cylinders is increased from 6 to 8.

Data Processing

First we need to load in the dataset and adjust the data frame based on the R Documentation.

```
data("mtcars")
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
```

```
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Analysis

This section is focused around making plots to see the correlation between some of the variables in the dataset. This will give us a visual representation of how certain variables in the dataset affect the MPGs.

Figure No. 1 in the Appendix is a pairs plot showing the comparison correlation of each variable. Visually, you can see that variables like number of cylinders (`cyl`), gross horsepower (`hp`), weight (`wt`), and the type of transmission (`am`) have a strong correlation with miles per gallon (`mpg`). When looking at the comparison between MPGs and the type of transmission, it appears manual cars have a higher MPG. Linear models will be used below to test this observation.

Figure No. 2 in the Appendix shows a simple boxplot comparing the miles per gallon to the type of transmission. In this plot, you can clearly see an increase in MPGs in cars with a manual transmission. Again, this will be tested with some linear regression models.

Figure No. 3 in the Appendix shows the effect transmission type has on MPGs given the number of cylinders the car has. As you can see, when the number of cylinders increase, the MPGs in manual transmission cars become more and more like the MPGs in automatic transmission cars.

In conclusion of the exploratory analysis, we see that MPGs differs by transmission type, but in addition, we can see that transmission type isn't the only variable that may or may not affect MPGs. A regression analysis is performed to test whether variables other than transmission type affect MPGs. The analysis will also answer the following questions:

- Is automatic or manual transmission better for MPG?
- What is the MPG difference between automatic and manual transmissions?

Regression Analysis

The focus of this section is to find the best fitting model for the set of data. To do this we will first look at the summary for the linear model where the outcome is miles per gallon (`mpg`) and the only predictor is type of transmission (`am`).

```
summary(lm(mpg ~ am, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

As you can see, the adjusted R^2 value is 0.34, which means only 34% of the variation is explained by this model. To get a better fitting model, next we look at the summary for the linear model where the outcome is miles per gallon (`mpg`) and every other variable is used as a predictor.

```
summary(lm(mpg ~ ., data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
```

```
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

From this model you can that the variation explained by the model has increased to 80%. That's better, but not the best. To find the best model, we will use the `step` method to build multiple regression models and select only the best variables to predict the outcome of miles per gallon (`mpg`).

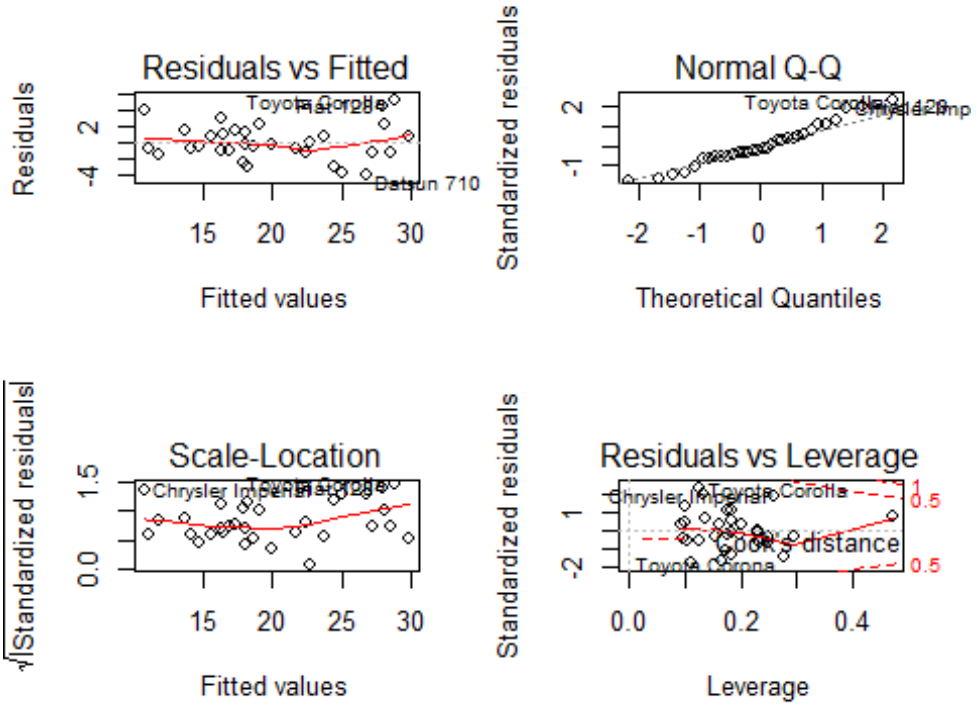
```
best <- step(lm(mpg ~ ., data = mtcars), direction = "both")

summary(best)

##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

From this model, you can finally see that 84% of the variation is explained, resulting in the best-fitting, multi-variable regression model. Now that we have the best-fitting model, we will plot the model residuals and diagnostics.

```
par(mfrow = c(2, 2))
plot(best)
```



From these plots you can see that Residuals vs. Fitted plot are randomly scattered which verifies the independence. The normal Q-Q plot has points that, for the most part, fall along the line which means that the residuals are normally distributed.

For statistical inference, a t.test was run to see whether the MPGs between automatic and manual car transmissions are the same. The results are as follows:

```
t.test(mpg ~ am, data = mtcars)

##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

From the results, the p-value is <0.05 which means we reject the null hypothesis in favor of the alternative. The true difference of the means is not equal to 0. In other words, the MPGs for automatic transmission cars is not equal to the MPGs for manual transmission cars.

Conclusion

Based on this regression analysis, we can conclude that manual transmission cars have a higher miles per gallon (`mpg`). After ajusting the miles per gallon for horsepower (`hp`), number of cylinders (`cyl`), and weight (`wt`), manual transmission cars get 1.8 more miles per gallon when comparing the automatic transmission cars. To further the analysis, we can conclude that MPGs will decrease by 2.5 for every 1000-pound increase in weight. MPGs will decrease 0.32 for every 10 increase in horsepower. When the number of cylinders increase from 4 to 6, the MPGs decrease by 3. And finally, the MPGs will decrease by another 2.2 when the numbers of cylinders is increased from 6 to 8.

Appendix

```
pairs(mtcars, main = "mtcars data")
```

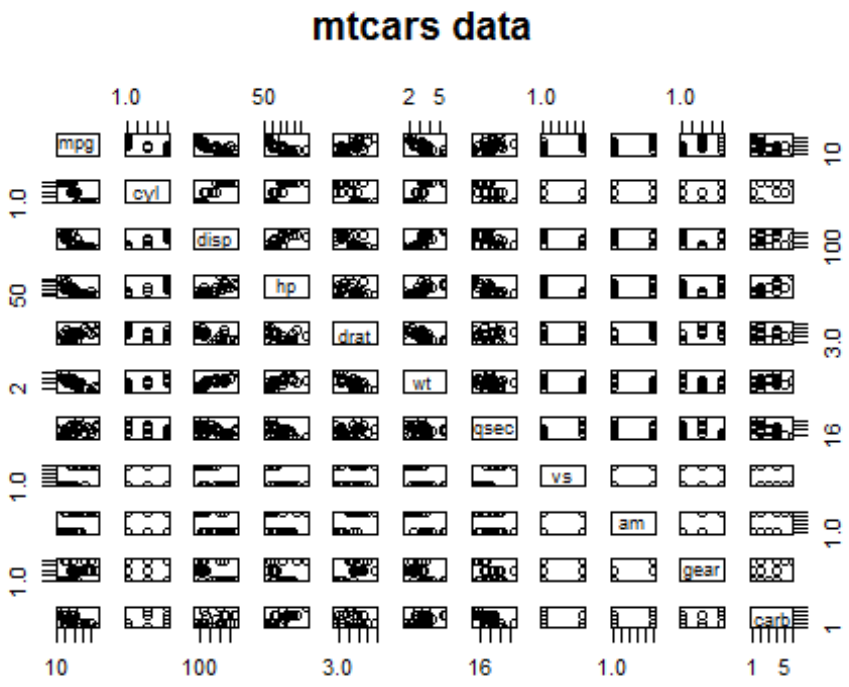


Figure No. 1 -- Variable Comarison

```r
boxplot(mpg ~ am, data = mtcars,
        col = c("red", "blue"),
        ylab = "Miles Per Gallon (mpg)",
        xlab = "Transmission Type (am)")
```

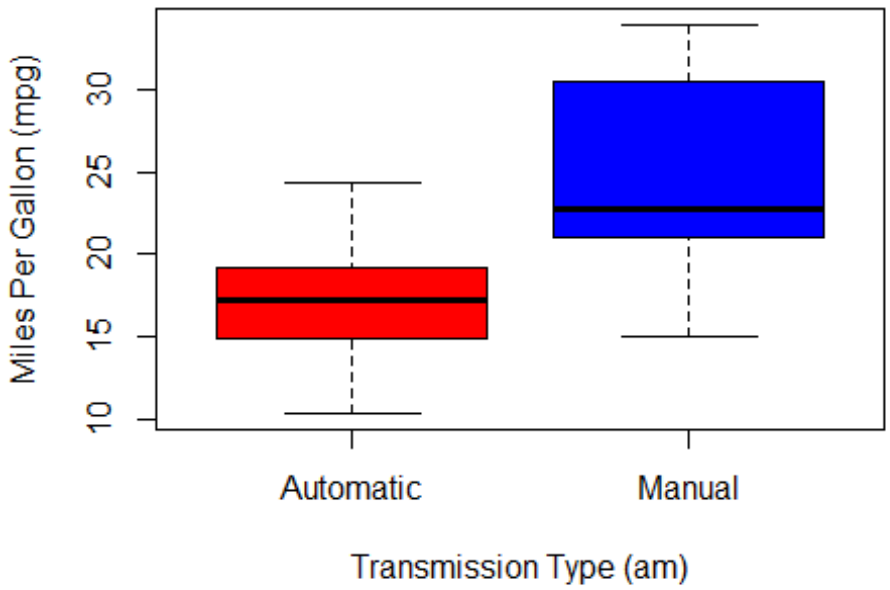


Figure No. 2 -- Effect of Transmission Type on MPGs

```r
coplot(mpg ~ am | cyl, data = mtcars,
       panel = panel.smooth, rows = 1)
```

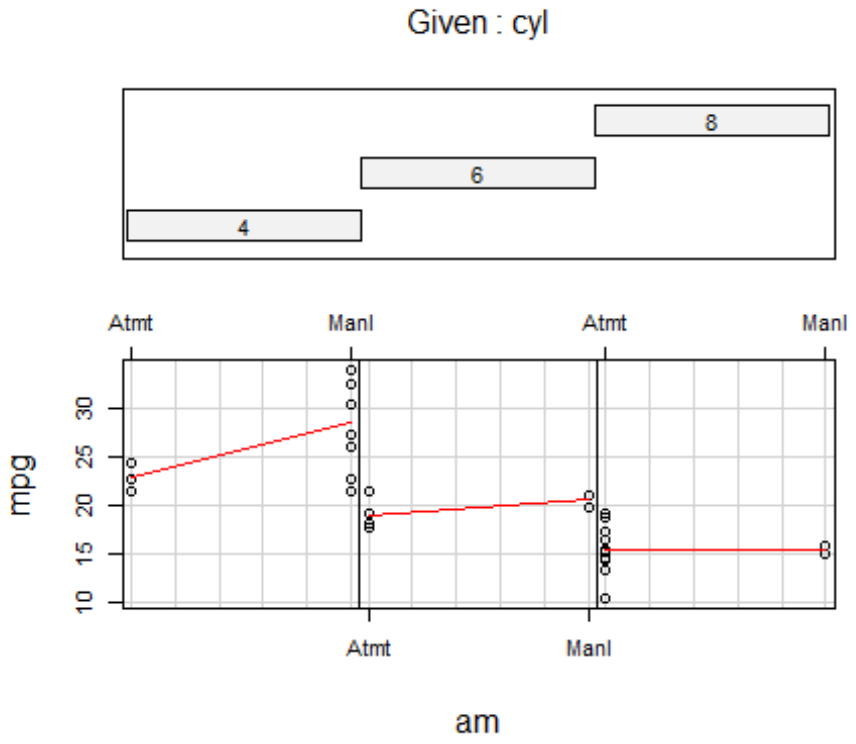


Figure No. 3 -- Effect of Transmission Type on MPGs Given the Number of Cylinders